The Integration of Similar Clinical Research Data Collection Instruments

Dorothy B. Cohen, MPH, Sandra J. Frawley, PhD, Mark A. Shifman, MD, PhD, Perry L. Miller, MD, PhD, Cynthia A. Brandt, MD, MPH
Yale Center for Medical Informatics, New Haven, CT

ABSTRACT: *We devised an algorithm for integrating similar clinical research data collection instruments to create a common measurement instrument. We tested this algorithm using questions from several similar surveys. We encountered differing levels of granularity among questions and responses across surveys resulting in either the loss of granularity or data. This algorithm may make survey integration more systematic and efficient.*

INTRODUCTION

The integration of similar clinical research instruments is a complex process that can benefit from informatics approaches that address issues such as data heterogeneity and semantic and granularity differences in questions and responses

BACKGROUND

The use of common measurement instruments can increase sample size and provide feedback on gaps in data collection on a single survey[1]. Previous work has been presented in the informatics literature on issues and methods to combine various standardized vocabularies[2]. In terminology mapping, important issues are content, nonvagueness, nonambiguity, and nonredundance of concepts[3]. These same attributes are also essential in developing a common data measurement. We will demonstrate one approach to combining multiple similar clinical research data collection instruments into one common instrument.

METHODS

We devised a simple algorithm for researchers to use for survey integration and applied our approach to 526 questions from several cancer risk factor surveys used by the National Cancer Institute (NCI) funded Cancer Genetic Network (CGN)[4].

We demonstrate the algorithm for the common concept "Ever smoked". Similar questions were, "Have you smoked at least 100 cigarettes in your lifetime? *(Yes, No)"*, "Have you ever smoked at least one cigarette a day for 3 months or longer? *(Yes, No)"*, " Ever smoked? *(Yes, No, Not App)"*, and "Have you smoked more than 100 cigarettes in your lifetime? *(No, Yes, Unk)"*. "Ever smoked" was selected as the Master. All other questions were mapped to this question and determined to be subset relationships. The Master response set was converted to, *"Yes, No, Not App, Unk"*. As the Master question does not define quantity or duration of smoking, granularity was lost from the three other questions.

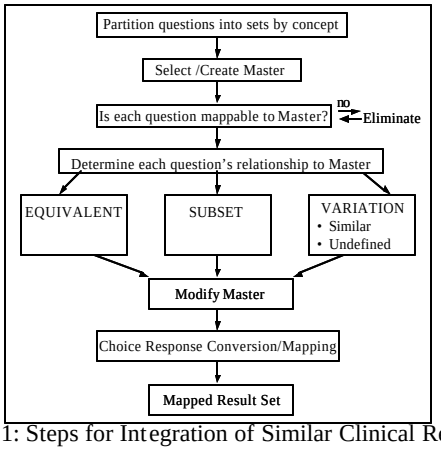

Figure 1: Steps for Integration of Similar Clinical Research Data Collection Instruments

RESULTS

The main issue we encountered was differences in granularity among questions and responses across surveys. In many instances there was more than one way to integrate questions, losing either granularity or data. The researchers' choice should be based on research design, sample size, and statistical methods.

CONCLUSION

The ideal time to use a structured approach to integration is prospectively in instrument creation. This type of algorithm can assist researchers in a more systematic and efficient approach to integrating similar clinical research data collection instruments for both prospective and retrospective integrations.

REFERENCES

1 Barrows, R. C., J. J. Cimino, et al. (1994). Mapping Clinically Useful Terminology to a Controlled Medical Vocabulary. SCAMC.

2 Dolin, R. H., S. M. Huff, et al. (1998). "Evaluation of a "lexically assign, logically refine" strategy for semi-automated integration of overlapping terminologies." Journal of the American Medical Informatics Association 5(2): 203-13.

3 Kannry, J. L., L Wright, et al. (1996). "Portability issues for a structured clinical vocabulary: mapping from Yale to the Columbia Medical Entities Dictionary." JAMIA 3: 66-78.

4 Cancer Genetics Network, National Cancer Institute, http://epi.grants.cancer.gov/CGN/.